VISION PAPER

Open Science and Data Science

Peter Wittenburg[†]

Max Planck Computing and Data Facility, Gießenbachstraße 2, 85748 Garching, Germany

Keywords: Open Science by Design; Open Science by Publication; Data Science; Data infrastructure; Digital Objects; FAIR

Citation: Wittenburg, P.: Open Science and Data Science. Data Intelligence 3(1), 95-105 (2021). doi: 10.1162/dint_a_00082

ABSTRACT

Data Science (DS) as defined by Jim Gray is an emerging paradigm in all research areas to help finding non-obvious patterns of relevance in large distributed data collections. "Open Science by Design" (OSD), i.e., making artefacts such as data, metadata, models, and algorithms available and re-usable to peers and beyond as early as possible, is a pre-requisite for a flourishing DS landscape. However, a few major aspects can be identified hampering a fast transition: (1) The classical "Open Science by Publication" (OSP) is not sufficient any longer since it serves different functions, leads to non-acceptable delays and is associated with high curation costs. Changing data lab practices towards OSD requires more fundamental changes than OSP. 2) The classical publication-oriented models for metrics, mainly informed by citations, will not work anymore since the roles of contributors are more difficult to assess and will often change, i.e., other ways for assigning incentives and recognition need to be found. (3) The huge investments in developing DS skills and capacities by some global companies and strong countries is leading to imbalances and fears by different stakeholders hampering the acceptance of Open Science (OS). (4) Finally, OSD will depend on the availability of a global infrastructure fostering an integrated and interoperable data domain—"one data-domain" as George Strawn calls it—which is still not visible due to differences about the technological key pillars. OS therefore is a need for DS, but it will take much more time to implement it than we may have expected.

1. INTRODUCTION

Advocating Open Science (OS) is important for reasons that have been enumerated frequently, such as the possibility to access the original data, upon which theories have been built [1]. This is especially

[†] Corresponding author: Peter Wittenburg (E-mail: peter.wittenburg@mpcdf.mpg.de; ORCID: 0000-0003-3538-0106).

important for the emerging Data Science (DS) paradigm as it was introduced by Jim Gray [2] since conclusions in DS are primarily based on data and algorithms operating on them. Science, of course, has many facets and uses a variety of paradigms. In this contribution, however, I will only focus on issues related to the DS paradigm and, due to its simplicity and great expressive power, I will make use of the concept of Digital Objects (DO), as defined by the Research Data Alliance (RDA)[①], which makes them first-class items on the Internet [3].

1.1 Putting Major Changes into Practice Takes Time

In discussions about OS people often extend the "act of classical publishing" to other scientific artifacts (data, software, workflow scrips, ontologies, etc.) [4], which indicates that the horizon of thinking is dominated by current practices. But this is not sufficient to fully understand the relevance of OS for the new DS paradigm and to implement the Open Science by Design (OSD) principle [5]. Instead, we need to anticipate future challenges and possibilities although this is not easy and might include assumptions that later may turn out to be wrong. The Internet pioneers keep telling us that when developing TCP/IP they had no idea how quickly it would be taken up or whether it might change the world. It took only three decades until people realized the new possibilities and began inventing innovative and revolutionary applications on top of the Internet, such as the Web. According to George Strawn, it took more than 200 years from the invention of the printing press in Europe until publishing scientific results became the norm [6]. The concept of the "semantic Web" (seamless semantic processing), introduced in 1999 by Tim Berners-Lee [7] has not yet affected most scientists' daily data work despite the development of many useful tools such as RDF (Resource Description Framework) and OWL (Web Ontology Language).

1.2 Observable Trends

There is no doubt that OS will become common practice in the DS domain, but at much slower speed than hoped, as George Strawn argues. One reason is that we still cannot anticipate which kind of structures will be established in the coming decades to revolutionize our data practices. We can only refer to some trends which are well-known: (1) Data volumes are increasing extremely implying that the individual DOs will lose relevance. We will base conclusions on specifically aggregated collections of DOs[②]. (2) The complexity of data will explode as well. It is not only the often-mentioned growing number of data types and formats which is summarized under the term complexity. It is the dramatically increasing number of

[①] Digital Objects have a content encoded in some structured bit-sequence which is stored in trustworthy repositories, and they are associated with metadata and have assigned a globally unique, persistent and resolvable identifier (PID).

[②] For an anecdote to illustrate this, I can refer to early discussions of how to best recognize speech. Phoneticians were dominating the discussions adding knowledge about the articulatory system and studying intensively individual speech sounds, their spectral representations, etc. They aggregated much knowledge and put this into rule-based systems. At a certain moment statistically based Hidden-Markov-Machines were introduced and proved to be much better in "characterizing" the essentials of the speech signal which then led to significantly better recognition results. The phoneticians who were looking at individual sounds and utterances were replaced by methods that extracted model parameters from large collections.

documented relationships between DOs that will increasingly represent essential parts of our scientific knowledge[3]. (3) It will become important to combine data from different silos in order to assist disciplines to come to new insights. Yet, cross-silo/disciplinary work is in its infancy since researchers in general do not seem to anticipate sufficiently well how cross-silo/disciplinary work could be efficiently supported and thus the hurdles are often too high [9]. (4) These trends clearly indicate that we need to move towards automated workflows for all kinds of processes in DS. At the end only such workflows based on simple and unifying core concepts such as DOs will help to make data FAIR which is why the concept of FAIR DOs was developed to make DOs fully FAIR compliant [10,11]. (5) Finally, as George Strawn argues, we will come to a scenario where we will see the overall data space as one coherent data space just as the Internet pioneers saw the network of computers as "one big computer". Yet, we do not know how this "integrated and interoperable data domain" will exactly look like and which software will effectively implement it. But given the expected huge cost savings it is obvious that experts will work hard on approaching this ideal.

1.3 Scope of OS

We also need to indicate what exactly is meant with OS in the DS domain to prevent misunderstandings and too high expectations. Obviously, we mean that most "relevant" DS artifacts should be exchanged, i.e., artifacts that have relevance for other work which includes later work in the same lab or work in other labs. Obviously, most of the raw data (after some screening, quality checking, etc.) qualify as being relevant. Even though some derived data and software will just be created to quickly test a procedure, for example, and thus have no further relevance to others[4], most of the derived data will qualify as relevant. The interpretation of what "relevant" means will remain fuzzy and depend largely on the opinions of the individual scientists and the evolving discipline cultures.

1.4 OS is More Than Accessibility

Currently almost all projects state that they intend to support the principles of FAIR and OS. We recognize, however, that these statements mostly indicate that projects will make their data findable and accessible in some way. This is already an important step ahead compared to the practices until now. OS in DS will only become effective when data will be "interoperable" and ready for efficient "reuse". This is one of the huge differences to "classical publishing" where texts are provided in natural languages knowing that our language capacity allows us to process the presented information, at least in a linguistic sense. In the DS domain we need to request that OS means that both, data and metadata (context and provenance) need to be available and machine actionable which implies still a long way to go. In DS, software is like a magnifying glass allowing us to inspect the content of data. Machine actionability enables scientists to easily do inspections and computations with the help of advanced software. Without this capability we cannot speak about OS in DS in my view.

[3] An excellent example of complexity can be found in the plans being worked out in the DISSCO initiative [8].
[4] This does not refer to work that does not deliver significant results which could also be relevant for others.

2. FACTORS HAMPERING OS

Having sketched some of the aspects of OS that are important in DS, I will now discuss six issues that may hamper fast progress in achieving OS, and that are often overlooked in the discussions.

2.1 Scope of Material

DS is driven by two major characteristics: (1) using novel statistical methods with as little priori knowledge included as possible not to bias the results; (2) aggregating large amounts of data to allow calculating the free parameters of the models. What is it we should make available for reuse in such a scenario?

- We could make the *algorithms* available, for example, for machine learning, but basic and highly optimized algorithms[5] have been widely studied and are anchored in public software libraries. Yet, a lab could create optimized versions or workflows combining specific methods which could be made available to others.
- We could make the *data collection available* which is used for training and which is crucial for obtaining the models. The selections incorporate much knowledge about the domain, and it is known that biased selections of data can lead to all kinds of results. The data may have undergone a variety of tests to prevent erroneous results.
- We could make the *model* available, i.e., the configuration and the set of parameters that have been calculated to achieve a certain result. In the example of artificial neural networks, however, we do not know, what the individual artificial neurons of the total ensemble are encoding and while the models may differ, they could exhibit a similar behavior. Different sorts of verifications and tests can be done to "understand" the behavior of the model and the results of these observations could be made available as well. Sharing models without the data that led to the models is not acceptable.
- In fact, by providing sufficient and correctly aggregated data collections and ready-made software everyone, including citizen scientists, could "press a button", create a model and use it for decision support, etc. Therefore, we can expect that a *large number of "models"* will be created for specific applications of which many can be executed on small computers either by fine-tuning existing basic models or creating new ones[6]. Which of these models will we want to make available and in which way? How should we evaluate these models?

[5] Researchers often work with complex models having millions of free parameters where a high degree of parallelization must be applied during learning, for example.

[6] It should be noted that some of these calculations require very large amounts of data and much compute power to calculate the model and that there is a trend to distributed model calculations due to the inclusion of sensitive data. In these cases, the software code needs to go to the data location.

Instead of classical publication including peer reviews, we will be faced in the DS domain with a huge validation challenge as should be evident from what has been described above. Models based on selected data collections and processed by a variety of algorithms will became available in great numbers to explain all kinds of phenomena or to enable all kinds of processing in all areas of science. A first principle must be that models without providing the references to the data collection being used and without offering its reusability must not be accepted. This alone might not be sufficient. It is not yet obvious what else needs to be done, but one way could be the request to register essential nano-publications, which are assertions augmented with key metadata, about the relevance of an offered model. This would allow users to run statistics about a variety of models and thus identify those that lead to erroneous results, outliers, etc.

2.2 Moment of Dissemination

We are facing a difference in views about when DS artifacts as discussed above should be made available to others. At the one extreme, publishers and some colleagues argue to repeat the current practices by waiting until a "classical publication" has been created, wrap up all needed data, software, etc., check their quality and then publish this package as well in a classical style. At the other extreme, many data scientists argue that data, in particular raw data, should be made available soon after creation[7] by simply making them DOs, i.e., self-standing "first-class items" on the Internet giving access to all its components via the associated permanent and resolvable identifiers. This dissemination act would be sufficient for reusing it since it would provide stable references and enable, for example, replications even after decades. The term "publication" is not used in this context on purpose, since it has connotations that come from centuries of classical publishing tradition, i.e., it is a slow process engaging intensive peer reviewing, cycles of improvements, etc. Waiting on a classical publication act would hamper fast innovation cycles and fast actions such as what is needed in the COVID case, which would not be accepted by the community.

Finally, DS will have much shorter cycles for exchanging most of its artifacts, but there might be a mixture of approaches. Three major reasons can be mentioned for short cycles: (1) We will see a dramatic increase of workflows being used in DS to improve FAIRness and efficiency and to unload the scientists from much administrative work, i.e., machine actionability enabling fast actions is key. (2) The sheer mass of scientific artifacts will increase dramatically, making it hardly possible to engage in a "classical publication" iteration. Making DOs available to others in a stable way will just be another easy step of the emerging canonical workflows. (3) Scientists want to immediately see what colleagues have done and want to immediately use brand-new data and software to work at the cutting edge. This is at least current practice for many data scientists.

For scientific reasons that go beyond the wish to replicate work, for example, most relevant components (software, data, workflow, model, etc.) should and will be made open at very early stages except that for some data limited embargo periods will be claimed. Surveys, however, indicate that researchers are not yet

[7] In many cases scientists argue for having an embargo period granted until making data available to others.

ready to change the procedures they are applying in the data labs [9]. Due to the ongoing dynamics in technology and the lack of proper software support they seem to be afraid of additional efforts to be taken without seeing the benefits. Using flexible workflow frameworks assisting scientists through all steps would help in making these artifacts available in a stable and trustworthy way during the process, thus implementing OSD. Despite the availability of many technical workflow frameworks, we lack scientifically motivated frameworks that guide the researchers through all steps without adding new complexities.

Summarizing, we can state that it will take quite some time until we understand how to effectively support researchers in their daily data practices and therefore increase acceptance for a change.

2.3 Roles in Data Creation and Recognition

It is argued that scientists will only accept OS, i.e., making data, software, and other scientific artifacts available to others, when there are solutions for appropriate incentives and recognition. To me this is not yet as simple as it may sound.

- As indicated, we should not simply transfer the classical "publication model" to DS, i.e., only small amounts of data will be published in the classical sense and the usual metrics cannot be applied.
- Another question that has been raised in discussions is whether the scientific value of, for example, creating a data collection has the same relevance as a scientific paper where theories are being discussed? Aggregating data and making it FAIR costs a lot of efforts and requires specific skills; however, data collections are just one of the preconditions enabling theorization. It is the smart selection and combination of data and algorithms based on deep domain knowledge that may lead to new insights. Indeed, acknowledging the different contributions is important, but their relevance in evaluations depends on job profiles and requires knowing the exact nature of contributions.
- Creating, maintaining, and offering a data collection for reuse in general requires the contributions of many actors with different backgrounds[®]. In many cases it is not a trivial undertaking to identify the exact role of the different contributors. Similar arguments hold for the creation and maintenance of software libraries that consist of many different contributions.
- Citations to data collections and software libraries can often not be done in the way we are used from classical frozen publications, since digital collections are dynamic (additions, deletions, and new versions) implying that the set of contributors will often change over time.
- What is today seen as a huge and complex task of data management will soon be taken over by highly standardized and canonical workflow frameworks that are populated by components that flexibly address specific needs. A recent analysis of more than 50 research infrastructures supported the assumptions of the RDA Data Fabric group [14] that data scientists use similar patterns of operations

[®] In the international DOBES project [12] a unique and nearly FAIR compliant data collection of about 20 TB was collected from about 70 teams worldwide including more than 200 scientists (linguists, biologists, ethnologist, etc.) and IT experts. In the NOMAD project [13] millions of data sets were aggregated from many labs worldwide, normalized and made FAIR compliant with a large number of contributors ranging from physicists to software developers.

in data management. It seems that the intelligence needed could be put into smart workflow software which would be yet another scientific artifact to be made available[9].

In times where roles are still subject of dynamic changes and where an agreed formal classification scheme of roles is missing, the exact contributions of individual persons are difficult to assess. Therefore, it will not be a trivial task to enable correct recognitions and give appropriate incentives that may have effects on career building, for example. Despite many ongoing efforts I would claim that we are not yet in a state to define appropriate metrics that will convince leadership of academic institutions to easily rate other artifacts than classical papers.

2.4 Global Competition

There is the usual competition between scientists to be the first with new findings and theories which, in general, is the main driver for their hard and inspired work, often in close collaborations. It implies that excellent scientists will protect the "cohort of artifacts" of their actual studies which includes all aspects from thoughts to methods and data at least for a limited time until it is obvious that no one else can pick up ideas. This attitude drove science to a large extent and the open exchange by classical publications was mostly the last step in this process of disseminating ideas and claiming ownership.

In addition to this "natural competition", we are now registering an increase in global competition with respect to exploiting large interdisciplinary data sets which is heated up by economic expectations ("data are the new oil" [15]). The major actors investing huge sums are some big globally acting information and publishing companies on the one hand and several countries and regions with sufficient financial capacity on the other hand. The ongoing huge investments will strengthen the already existing imbalances in the capability of exploiting data, especially in cross-disciplinary research. The expected advantages will be manifested in the knowledge of experts[10] and in the availability of an efficient and powerful infrastructure. The consequences of these investments can be counterproductive for OS, since every scientist will immediately understand that Open Data implies that the strong players will use the scientists' data to come up with new findings and, where economic interests are involved, to make money. Fears can be observed not only in the academic world:

- Scientists are afraid that their data will be exploited by others and that their role is degraded to mere data providers.
- Scientists and funders are afraid that costly services will be established on top of their data for which they have to pay.

[9] To indicate the trends we can refer to the example of a large US university collaboration which hired two FTE for three years to create such a generic workflow template to make cancer research much more efficient and reproducible.

[10] The capability of educating data scientists and data managers, providing excellent conditions and paying high salaries will give advantages. Even worse is that, following current trends, it seems that the imbalances between countries and companies will increase.

- Governments, funders and research organizations are afraid that their investments will not pay back.

It would be naive to ignore these fears. Currently they are influencing discussions and the willingness to make data open, although in the public other words may be chosen. Global competition inspired by commercial and national interests is already hampering OS. There is no doubt that companies will have an important role to advance infrastructure and tool building as they had for establishing the Internet, but we should note that for making the Internet new types of companies were finally needed to achieve the great breakthroughs. Yet, we seem not to have an idea which kind of companies would effectively help researchers to establish an OS domain. It will take a while until new mechanisms to guarantee data sovereignty will have been settled that help to overcome fears.

2.5 Unbalance in Skills

A highly related aspect has to do with an increasing imbalance in skills. In 2000 we started the global DOBES project to document endangered languages with the intention to make as much data as open as possible for others [12]. This was a revolutionary act in linguistics, since then the domain experts, who were used to only refer to papers and thus to the analysis results of their colleagues, were able to work on the data[⑩] directly. After many discussions in the project an "open as default" policy was agreed upon. One of the most prominent factors that hampered the willingness of the linguists to make data open was the fear that only a few colleagues had (and still have) the skills to carry out studies based on data from different scientists. This had not only to do with the lack of technical skills, but also with the lack of imagination how this paradigm shift could be used to create results that would be respected by the scientific community. How could one carry out, for example, a comparative study on intonation patterns in different languages using the speech waves directly? Currently, only few researchers in the domain would have the skills to do this kind of work.

This line of argumentation can be extended to the huge differences in skills and facilities between industrial and developing countries. Data being generated in developing countries could be used and exploited in countries and by companies that have much better conditions. A phase of "data colonization" might occur which necessarily would lead to closures of data.

Therefore, we also must cope with factors related to generations and cultures, factors which change only very slowly. It will take much time to educate a new generation of scientists in many disciplines across countries who understand the new paradigm, who acquainted the required skills and who dare to undertake new steps in research which in a ramp-up phase may not get the recognition in the classical sense. And there is a big task to carry out much more training of new generations of experts in developing countries.

⑩ These were mainly media recordings, layers of different linguistic annotations, metadata and other linguistic products such as lexicons and grammars.

2.6 Data/Research Infrastructure

There is no doubt that the availability of a proper "data/research infrastructure" will be decisive for an efficient DS. This has many aspects—of which I will mention just one. It is known that about 80% of the costs in data projects are due to "data wrangling" which summarizes all steps before one finally can start the analytics [16]. If scientists have an infrastructure at hand which would reduce the waste of time by a half, for example, an enormous cost reduction and an increase in efficiency would be possible.

Currently, regions such as Europe and countries such as Germany spend about 90 M € per year on improving data infrastructures. The big companies spend billions of dollars in improving their data infrastructures, which extend from fast cloud systems to improving syntactic and semantic processing and statistically based analytics.

Yet, we do not know exactly what the key pillars of such a data infrastructure reducing the inefficiencies considerably will be: (1) Big companies that will try to create dependencies and impress their way of doing on everyone. This strategy will finally fail due to the international competition. (2) Those who sell the cloud-systems as THE solutions which is naive since the major inefficiencies are caused by deficits in data organization and in syntactic and semantic explicitness. (3) Those who believe that by improving the currently available services landscape stepwise will lead to the solution. This will fail as well, since finally services will populate an infrastructure but not form the core of a FAIR compliant approach. (4) Those, including the author, who believe that a fundamental new orientation based on neutral suggestions such as indicated with FAIR DOs is needed to create a momentum comparable to what TCP/IP achieved in networking a few decades ago. As usual a mixture of approaches will evolve, but it needs to be based on global agreements since otherwise OS will be severely hampered.

Despite the successful activities of some global initiatives such as W3C, RDA, CODATA and DONA to organize the discussions around standardized pillars for an efficiently working data infrastructure, we do not yet seem ready for global agreements and breakthroughs. Too many different voices and interests on the one hand and missing conviction about suggested open solutions such as the FAIR DOs on the other hand hamper fast progress. The FAIR principles were a great step in achieving convergence but are not sufficient to determine the pillars of a future data infrastructure. Finding global convergence on infrastructure pillars will obviously take some time to evolve.

3. CONCLUSIONS

The DS paradigm is dependent on a high degree of findability, accessibility, interoperability, and reusability (FAIR) of data, software and other artifacts across silos and disciplines. It is claimed that DS will necessarily be dominated increasingly often by automatic workflows which implies that the dimensions of interoperability and reusability will become crucial to make DS effectively happen. Machines need to be able to work on data without human intervention, since otherwise we will hardly be able to reduce the huge inefficiencies and costs in data work and thus to make DS effectively be used by many.

Of course, there will be artifacts that will not be shared, but at least a large fraction of data being aggregated as a result of experiments, observations, simulations and computations will be relevant or has the potential of becoming relevant for further scientific work. This requires that when discussing the act of making the DS artifacts available in a persistent and stable way supporting FAIR principles, we need to add new smart processes in addition to the classical publication processes which evolved over hundreds of years. Only with these new processes, which will include smart validation options for the many DS products, we will effectively enable OS in the DS domain. These emerging processes can be characterized as

- being executed by components of automatic workflows;
- encapsulating all relevant information about an artifact in a stable and machine actionable bundle which we call FAIR DO; and
- providing DS artifacts for reuse at an early stage.

Yet, we lack clear and accepted ideas how these processes will exactly look like and which structures and mechanisms will be required to support OS in the DS domain. Many fundamental aspects such as the possible usefulness of the FDO concept are still debated. Currently there are not sufficient actors who understand that simplification of some underlying structures and mechanisms is the way to cope with the increasing complexity. Mechanisms imposed by big companies will fail to create globally accepted standards and thus to come to breakthroughs since sensitive national and economic interests will be touched. Obviously, we need new types of companies supporting our ideas of OS in the same way as new companies supported the open Internet movement.

Major obstacles for establishing an effective domain of OS are not of technological but of cultural nature. It seems that a new generation, which is well-trained in DS methods, will be required to create a world where data providers are part of the exploitation work and not just second-order contributors. National investments in data/research infrastructures are important to improve the conditions for OS, but we should not overlook that many countries cannot participate adequately and that we therefore need measures to compensate for the increasing imbalances.

Roles in the creation process of DS artifacts will still change due to the dynamic developments. We are not ready yet to describe them sufficiently well, making it hard for academic hierarchies to judge the nature and relevance of individual contributions for recruitment purposes.

The overall conclusion therefore is that the cultural situation exhibits essential gaps and that essential technological components are missing that prevent us to quickly establish an effective OS domain. We can see broad initiatives such as RDA, CODATA, GO FAIR, and GEDE that are working on these aspects and we also see an increasing number of projects testing out components that point into the right directions; however, we obviously will need more time to work out adequate OS structures, mechanisms and technologies. Of course, all excellent efforts from librarians, researchers and other experts being currently engaged in pushing the OS agenda need to be continued.

REFERENCES

[1] Burgelman, J.-C., et al.: Open science, open data, and open scholarship: European policies to make science fit for the twenty-first century. Frontiers in Big Data 2,43 (2019)

[2] Hey, T., Tansley, S., Tolle, K. (eds.): The fourth paradigm: Data-intensive scientific discovery. Microsoft Research, Redmond (2009)

[3] RDA DFT core terms and model. Available at: http://hdl.handle.net/11304/5d760a3e-991d-11e5-9bb4-2b0aad496318. Accessed 5 January 2021

[4] Open science. Available at: https://en.wikipedia.org/wiki/Open_science. Accessed 5 January 2021

[5] National Academies of Sciences, Engineering, and Medicine. Open science by design: Realizing a vision for 21st century research. The National Academies Press, Washington (2018)

[6] Strawn, G.: Open science and the hype cycle. Data Intelligence 3(1), 88–94 (2021)

[7] Semantic web. Available at: https://en.wikipedia.org/wiki/Semantic_Web. Accessed 5 January 2021

[8] GEDE-RDA-Europe/GEDE. Available at: https://github.com/GEDE-RDA-Europe/GEDE/blob/master/Digital-Objects/DO-Workshops/Workshop-Philadelphia-2019/koureas-do-p13.pdf. Accessed 5 January 2021

[9] Jeffery, K., et al.: Not ready for convergence in data infrastructures. Data Intelligence 3(1), 116–135 (2021)

[10] Paris-FDO-workshop. Available at: https://github.com/GEDE-RDA-Europe/GEDE/tree/master/FAIR%20Digital%20Objects/Paris-FDO-workshop. Accessed 5 January 2021

[11] Bonino at GEDE Paris Session. Available at: https://github.com/GEDE-RDA-Europe/GEDE/blob/master/FAIR%20Digital%20Objects/Paris-FDO-workshop/GEDE_Paris_Session%201_Bonino.pptx. Accessed 5 January 2021

[12] DOBES. Available at: https://dobes.mpi.nl/data. Accessed 5 January 2021

[13] NOMAD Centre of Excellence. Available at: https://nomad-coe.eu/. Accessed 5 January 2021

[14] Data Fabric IG. Available at: https://www.rd-alliance.org/group/data-fabric-ig.html. Accessed 5 January 2021

[15] Data is the new oil of the digital economy. Available at: https://www.wired.com/insights/2014/07/data-new-oil-digital-economy/. Accessed 5 January 2021

[16] Wittenburg, P., Strawn, G.: Common patterns in revolutionary infrastructures and data. Available at: http://doi.org/10.23728/b2share.4e8ac36c0dd343da81fd9e83e72805a0. Accessed 5 January 2021

AUTHOR BIOGRAPHY

Peter Wittenburg was Executive Director of Research Data Alliance (RDA) Europe, Member of RDA Technical Advisory Board, and Scientific Coordinator of European Data Infrastructure (EUDAT). He set up and led the Technical Group with about 30 experts at Max Planck Institute (MPI) for Psycholinguistics and then led the Language Archiving Group with about 25 experts. Since 2000 he has played leading roles in a variety of European (funded by the European Commission) and national projects (funded by MPS, DFG, BMBF, NWO 23) and ISO initiatives (ISO TC37/SC4). He won the Heinz Billing Award of the Max Planck Society (MPS) for the advancement of scientific computation in 2011 and received an honorary doctorate from University Tübingen in 2013.

ORCID: 0000-0003-3538-0106